共起情報を考慮した $tf \cdot idf$ 法に基づく関連文書間の自動ハイパーテキスト化

岡村 潤　大森 信行　山口 登志実　森 辰則　中川 裕志
横浜国立大学 工学部 電子情報工学科
E-Mail: {jun,ohmori,yamaguti,mori}@forest.dnj.ynu.ac.jp, nakagawa@naklab.dnj.ynu.ac.jp

概要

今日，高機能・高性能なシステムに付属されるマニュアルは別冊に分割されていることが多く，利用者が必要とする情報を効率よく得ることは困難となっている．本論文では，複数マニュアル間の関連箇所を自動的に抽出しハイパーリンクを生成する方法について述べる．本研究では，マニュアル内の段落を1つの文章のまとまりと考え，段落中のキーワードに $tf \cdot idf$ 法によって重みをつけ，ベクトル空間法により段落間の類似度の大きさを求め，その関連度にしたがって段落間でのハイパーリンクを自動生成する．また，より高精度な関連付けのために段落間でのキーワードの共起情報とキーワードの語彙連鎖を用いることが有効であることが確かめられた．

1 はじめに

今日，各種機器やソフトウェアが高機能・高性能になるにつれて，それに付属されるマニュアルも利用者のレベルや使用用途別に合わせて，複数に分冊されるような形態をとるようになってきた．これに伴いユーザもそれらのマニュアル群からユーザ自身が欲する知識・概念を取り出すことが必要になってきている．そのような必要な知識・概念を得るのに，従来の紙面のマニュアルにおいては参照すべき部分を目次や索引から探しだし，読み進めていくことによってなされていた．しかしながら複数に分かれたマニュアル群において，目次や索引から参照箇所を探しだす作業はユーザに大きな負担をかけることになる．

このような負担を軽減するために，最近では文書や図の間の相互参照情報(ハイパーリンク)をもつハイパーテキスト型マニュアルが見受けられるようになってきた．しかし，このハイパーリンクの構築はあらかじめ人間の手作業によって行なわれる必要があり，大規模マニュアル群におけるこの作業は困難を極める．本稿では，複数マニュアル間におけるハイパーリンクを情報検索手法を用いて自動的に生成するシステムを提案する．マニュアルにおいては語句の説明箇所の他に，一連の操作手順などが書かれたまとまった文書が参照対象になり得る．例えば，初心者用マニュアルに例示されている操作について，それに対応する詳細記述をリファレンスマニュアルで調べる場合などが想定される．そこで本稿では，図1に示すような節や項などのまとまった文書単位での自動ハイパーリンク生成を考える．
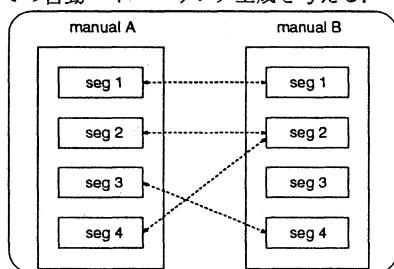
図1: システムが生成するハイパーテキストの概念図

2 自動ハイパーテキスト生成システム

本システムにおいて，我々は自動ハイパーテキスト生成を次のように実現する．

1. ハイパーリンク生成の対象は，文書小単位(セグメント)である．2つのマニュアルをセグメント単位に区切り，その全てを候補と考える．セグメントの単位としては，意味的なまとまりを考慮し，節，項とする．

2. 関連付けについては，まず両マニュアルにおける任意の組合せについてあらかじめ類似度計算を行っておく．ハイパーリンクは利用者に提示する時に，類似度の高いものから動的に生成し，提示する．

1. については，HTML など構造をもつ記述形式になっていれば，文書構造からセグメントを認識できるため容易に自動化できる．

2. については，類似度のスコア付けが問題となる．この類似度のスコア付けには，情報検索で広く用いられている，$tf \cdot idf$ 法に基づくベクトル空間モデルを利用する．

2.1 システム概要

本システムは，図2に示す4つのサブシステムより構成されている．

本システムでは，以下のような手法でセグメント間類似度を計算する．

1. 対象となる文章群から語を抽出し，それぞれに $tf \cdot idf$ 値によって重要度を与える．

2. ベクトル空間モデルによってセグメント間の類似度を計算する．類似度計算においては，以下のようなベクトルを生成する．
   - 一つのセグメントに一つのベクトルを対応させる．
   - ベクトルの各次元には各単語が，各成分には対応する単語の重要度，すなわち単語の $tf \cdot idf$ 値が割り当てられる．
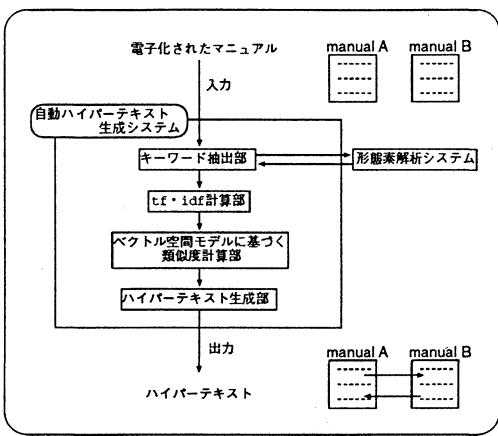
図 2: 自動ハイパーリンク生成システムの構成

システムの利用画面の例を図3に示す．画面を縦分割し，左右フレームに異なる2つのマニュアルがそれぞれ表示される．それぞれのフレームは更に上下分割され，現在表示されているセグメントのリンク先が表示されており，いずれかをクリックすることにより，参照先がそれぞれのフレーム上部分に再表示される．その後も同様にリンク先をたどっていくことができる．
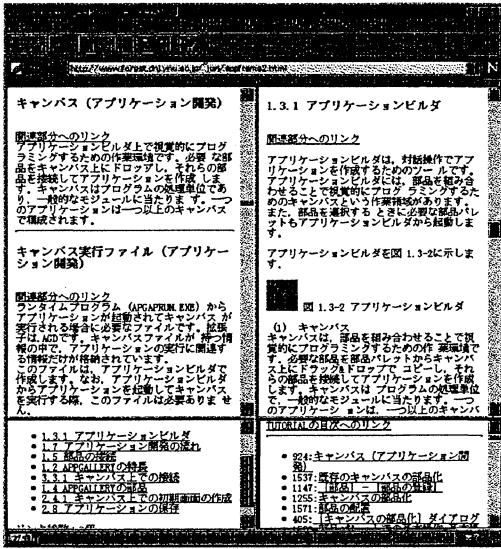
図 3: システムの利用画面

3 より高精度な関連付けのために

本システムでは，文書中の操作の対応に基づいてセグメント同士を対応付けることを目的としている．さらなる高精度なセグメント間の対応付けのために，我々は

1. 格情報・共起情報
2. 語彙連鎖 (Lexical Chain)

を関連度計算に反映させることを考えた．

3.1 格情報・共起情報の利用

一般的に，操作の説明は「スイッチをビデオ側に合わせる」のように

名詞 1- 格助詞 1 名詞 2- 格助詞 2 … 動詞

といった操作対象を表す名詞と操作内容を表す動詞で表される．そこで，文中の名詞や動詞の間の関係を利用することで，その文の表す操作 (の一部) に重きを置くことにより，より高精度なセグメント間の対応を取ることができると考えられる．例えば，2 セグメント内の文に同じ名詞対が共起した場合にはセグメント間の類似度に共起した名詞の重要度に応じた値を加算し補正を行うことなどが考えられる．我々は，単語の共起情報によってセグメント内の単語頻度 tf を補正して類似度計算を行う．

3.1.1 共起情報を単語頻度 tf を補正する方法

情報検索における文書の重要度決定に，検索要求文内で共起している単語対の共起重要度を利用すると，同じ再現率に対する適合率が向上することが報告されている [高木 96]．本稿では文書間のハイパーテキスト化を考えているので，対象となる両方のマニュアルについて，出現する全ての共起単語対についての共起重要度 cw を計算し，類似度計算に反映させることを考える．さらに高木らの方法に加えて格情報を考慮する．

図 4に本手法のイメージを示す．
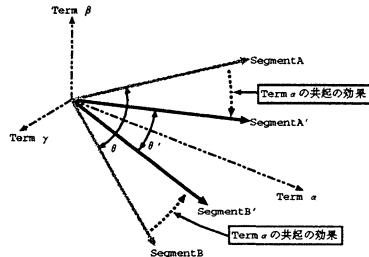
図 4: 共起を類似度計算に影響させるイメージ

共起による効果は vector 計算をする時に付加する．実際には，共起は単語の対を単位としているので，本手法では操作としての共起があるところに対して影響をあたえていることになる．

この手法では，2 セグメント d_A, d_B 間の類似度計算において，両セグメントに出現している共起単語対について，tf の値を次のように補正する．ある語 t_k がセグメント d_A に f 回出現した場合，新たに $tf'(d_A, t_k)$ を文書内出現頻度として語の重要度を算出する．$tf'(d_A, t_k)$ は以下の式により計算する．

$$
\begin{aligned}
tf'(d_A, t_k) &= tf(d_A, t_k) \\
&\quad + \sum_{t_c \in T_c(t_k, d_A, d_B)} \sum_{p=1}^{f} cw(d_A, t_k, p, t_c) \\
&\quad + \sum_{t_c \in T_c(t_k, d_A, d_B)} \sum_{p=1}^{f} cw'(d_A, t_k, p, t_c)
\end{aligned}
$$

ここで，$T_c(t_k, d_A, d_B)$ は d_A, d_B の両セグメントで t_k とある範囲内の位置で共起している単語の集

合である．p は，セグメント d_A 内で，ある語 t_k が出現する場所を表しており，セグメント内の全ての出現箇所に対しての cw の和を計算している．この計算を T_c に含まれる全ての単語について行い，tf に加算する値を得る．

また，cw は，共起を調べる単語として名詞のみを考慮した共起重要度であるが，cw' は，名詞とその直後に出現する格助詞を一つの単語と考え，cw と同様に求めたものであり，格助詞と名詞の組に関する共起に着目した共起重要度である．

次に共起重要度 cw の算出法を説明する．cw' についても名詞と格助詞の組を1 つの単語と見なす以外は算出法は同様である．まず，t_k と t_c における語間の近接出現係数 $\alpha(d_A, t_k, p, t_c)$ と共起係数 $\beta(t_k, t_c)$ を次のように定義する．

$$\alpha(d_A, t_k, p, t_c) = \frac{d(d_A, t_k, p) - dist(d_A, t_k, p, t_c)}{d(d_A, t_k, p)}$$

$$\beta(t_k, t_c) = \frac{rtf(t_k, t_c)}{atf(t_k)}$$

$d(d_A, t_k, p)$ はどれくらいの距離までを共起の範囲とするかを表すパラメタである．本稿では 1 つの意味的なまとまりである一文中の単語の共起を見ており，$\alpha(d_A, t_k, p, t_c)$ は文内に共起した単語についてのみ計算する．よって，$d(d_A, t_k, p)$ は注目している動詞句が出現している一文の単語数である．また，$dist(d_A, t_k, p, t_c)$ は，セグメント d_A で p 回めに出現した t_k について単語数で計算した t_c との距離である．$atf(t_k)$ は注目しているマニュアル内の t_k の出現総数，$rtf(t_k, t_c)$ は一文内に共起している t_k と t_c の出現総数である．

次に，t_k の共起語 t_c の近接出現共起単語の重要度 $\gamma(t_k, t_c)$ を定義する．N は各マニュアル中のセグメント数であり，$df(t_c)$ は t_c の出現する文書数である．

$$\gamma(t_k, t_c) = log(\frac{N}{df(t_c)})$$

以上で定義した，近接出現係数 $\alpha(d_A, t_k, p, t_c)$，共起係数 $\beta(t_k, t_c)$，接出現共起単語重要度 $\gamma(t_k, t_c)$ から，セグメント d_A 内の p 番めに出現する

語 t_k の共起重要度を次の式で表す．

$$cw(d_A, t_k, p, t_c) = \frac{\alpha(d_A, t_k, p, t_c) \times \beta(t_k, t_c) \times \gamma(t_k, t_c) \times C}{M(d_A)}$$

$M(d_A)$ はセグメント d_A 内の形態素数であり，tf と同様の正規化を行なっている．C は共起重要度正規化係数である．この値は，大きいほど共起重要度が tf にあたえる影響が大きくなる．

3.2 語彙連鎖の利用

一般にマニュアル内容にも話の流れがあり，$tf \cdot idf$ 法では検出されないような，複数のセグメントにまたがって出現する重要な概念がたびたび登場する．

このような概念をとらえることができれば，セグメント間の対応付けの精度向上に有用であると考えられる．我々は，この効果を類似度計算に反映させるために語彙連鎖の導入を考えた．

3.2.1 語彙連鎖

語彙連鎖とは，文章中で語彙結束関係にある語のまとまりのことである．一般に言う語彙結束性とは，語の意味的なつながりのことである．この意味での語彙連鎖は概念辞書上の同一カテゴリ (意味分類) に属するものとして計算される [Gre96]．しかし我々のシステムは同カテゴリのマニュアル群を対象としているため，同じ単語は同語義を示し，また同じ概念は同じ表記の語により指し示されることが期待できるため，シソーラスなどの概念辞書を用いる必要がないと期待できる．そこで，本稿における語彙連鎖とは「文書中で同じ語が連続して出現している部分」のことを指すことにする．

3.2.2 語彙連鎖の効果と導入

語彙連鎖を用いると，セグメントを越えて出現する語を捉えることにより，複数のセグメントに渡る重要語を見つけることが可能となる．逆に，操作名・機能名が羅列されているセグメントは，語彙連鎖が構成されにくいため，対応の重要度を相対的に下げる可能性もある．

本稿における語彙連鎖の概念図を図 5に示す．
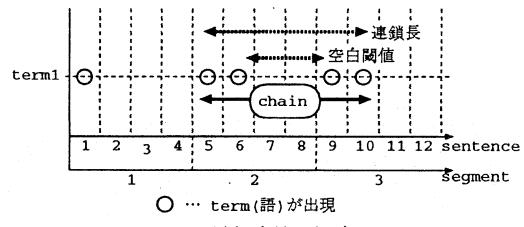
図 5: 語彙連鎖の概念

我々のシステムにおける語彙連鎖は，語彙結束関係を意味的なつながりでなく，同一語の連続出現性とみなしている．よって，語彙連鎖はセグメント軸上に構成されることになる．

例えば図5において，セグメントをさらにいくつかの単位 (ここでは最小セグメントと呼ぶ．図 5においては最小セグメントは一文) に分割し，その最小セグメント中に，注目した語 (図 5の例では "term1") が出現するかどうかをチェックする．図 5中の "○" がこの "term1" が出現したことを示している．

そして，連続して語が出現している (図5において "○" が連鎖を形成している) 場所において語彙連鎖を構築する．

このような処理を，セグメント間類似度を計算するために抽出された語すべてについて行ない，それぞれの語について語彙連鎖を構築する．

ある語が語彙連鎖を形成している場所は，その語についての何らかの説明・操作内容などが記述されていると考えることができる．

よって，語彙連鎖により同じ語でも話題の中心になっている時とそうでない時の重要度の差をつけることができる．

語彙連鎖の導入においては，

- **連鎖長閾値:** 連鎖の長さに関するパラメタ．ここでは，以下の２つを考える．

- 長連鎖閾値: どの長さ“以下”の連鎖を語彙連鎖と見なすか. 逆に, どの長さ以上の語の連鎖は語彙連鎖とみなさないか.
- 短連鎖閾値: どの長さ“以上”の連鎖を語彙連鎖と見なすか. 逆に, どの長さ以下の語の連鎖は語彙連鎖とみなさないか.

- 空白閾値: 語と語の間の出現距離 (時間的距離), すなわち語の間のギャップ (空白) をどこまで無視し, 連鎖とみなすか.

という3つのパラメタが問題となる. 今回の実験ではこれらのパラメタは平均適合率を最大にするものを選んだ.

語彙連鎖をセグメントの類似度計算に反映させるには, 語の tf 値を補正を行なうことにした. すなわち, あるセグメントにおいて語彙連鎖を構成している語に対してその語の tf 値を以下のように補正する.

$$tf'(d_i, t_k) = tf(d_i, t_k) \times (1 + \alpha)$$
$$(\text{ただし}\alpha > 0)$$

α については, 平均適合率を最大にするものとして今回は $\alpha = 3.5$ としている.

4 システム評価

評価には, 情報検索で一般的に利用される再現率 (*recall*), 適合率 (*precision*) を用いる.

$$\text{再現率}(recall) = \frac{\text{検索された適合対応数}}{\text{全ての適合対応数}}$$
$$\text{適合率}(precision) = \frac{\text{検索された適合対応数}}{\text{検索された対応数}}$$

再現率はある順位までに出現する正解の割合, 適合率はノイズの割合をそれぞれ示す.

我々は, 2で述べてきたシステムに対して, 大規模マニュアルを対象とした評価を既に行なっており, そこで, ある程度のシステムの有効性を確認している [大森 97].

しかし大規模マニュアルにおいては完全な正解集合を人手で作るのが困難であるため, 3で述べた手法の効果を考察するには, 完全な正解集合を作成することができる小規模のマニュアルで実験を行う必要がある. そこで我々は, 同一メーカの 2 つのビデオのマニュアル [三菱電 a, 三菱電 b] 間で本システムによるハイパーテキスト化を行い, 対応の評価を行なった. 各マニュアルのセグメント数と大きさは, マニュアル A [三菱電 a] が 32 (80 kbyte), マニュアル B [三菱電 b] が 28 (98 kbyte) である. 全セグメント数対応 896 のうち, 正解は 60 通りであった. 今回は

1. $tf \cdot idf$ 法のみによる語の重み付け
2. 共起情報を考慮した語の重み付け
3. 語彙連鎖を考慮した語の重み付け

という三つの手法によるセグメント対応の結果を比較の対象とした. 今回, 語彙連鎖についてはマニュアル別に各パラメタを設定した.
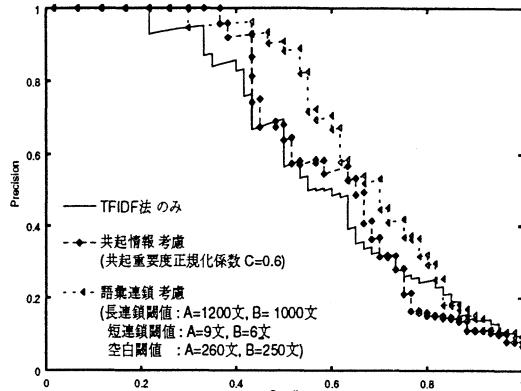
図 6: 適合率と再現率グラフ

適合率・再現率グラフを図 6に示す.

図 6より, 3で述べた二つの手法が, $tf \cdot idf$ 法のみによる手法より, 再現率の低～中域で対応の精度が向上していることが分かる. また語彙連鎖については, 複合名詞を考慮しないほうが高精度なセグメント対応を得ることができた.

5 おわりに

本稿では, 複数のマニュアル間における自動ハイパーテキスト化手法について述べ, さらにセグメント対応の精度を向上させるための二つの手法を提案し, 評価実験を行い, その有効性を示した.

ただし, 語彙連鎖については各パラメタとマニュアル構造の関係特性について, 更に実験を重ねて調査する必要があるであろう.

また, 今回は各手法のみの効果を知るために独立した実験を行なっており, お互いを組み合わることはしていない. よって, この二つの手法を組み合わせることによるセグメント対応の精度向上の有無を検証してみる必要がある.

参考文献

[Gre96] Stephen J. Green. Using lexcal chains to build hypertext links in newspaper articles. In *Proceedings of AAAI Workshop on Knowledge Discovery in Databases, Portland, Oregon*, 1996.

[大森 97] 大森信行, 蔵方隆宏, 岡村潤, 森辰則, 中川裕志. 情報検索手法を用いた複数文書間の関連箇所抽出 – 電子化マニュアルへの適用 –. 言語処理学会第 3 回年次大会, pp. 257–260, 1997.

[高木 96] 高木徹, 木谷強. 単語共起関係を用いた文書重要度付与の検討. 情報学基礎研究会報告 96-FI-41-8, 情報処理学会, 1996.

[三菱電 a] 三菱電機株式会社. 三菱ビデオ HV-FZ62 取扱説明書.

[三菱電 b] 三菱電機株式会社. 三菱ビデオ HV-F93 取扱説明書.